УДК 311.17

Єріна А. М.

ПОЄДНАННЯ МОДЕЛЕЙ РЕГРЕСІЙНОГО І ДИСПЕРСІЙНОГО АНАЛІЗУ В СОЦІАЛЬНО-ЕКОНОМІЧНИХ ДОСЛІДЖЕННЯХ

Розглядаються аналітичні можливості та варіанти адаптації, розширеної регресії до неоднорідних сукупностей і згрупованих даних.

Соціально-економічні явища масові, надзвичайно складні, взаємопов'язані і взаємозумовлені. Закономірності формування та розвитку будь-якого з них характеризуються багатовимірною системою факторів. Вони недоступні прямому спостереженню і не піддаються експериментуванню; характерними їх рисами є випадковість та невизначеність. Ефективним засобом пізнання таких закономірностей є статистичні моделі, зокрема, при вивченні взаємозв'язків — моделі регресійного та дисперсійного аналізу, які й за умов невизначеності забезпечують сталість і надійність висновків щодо сили осібного й сукупного впливу факторів.

Обидві моделі ґрунтуються на вивченні спряженої варіації двох чи більше ознак взаємозв'язаних явищ. Основне завдання дисперсійного аналізу — виявити джерела варіації; дисперсійні комплекси орієнтовані переважно на обробку згрупованих даних з однаковою частотою груп і підгруп. У соціально-економічних дослідженнях групування є результатом статистичного спостереження і зведення даних, а, отже, забезпечити однакові частоти розподілів практично неможливо. А тому перевага віддається регресійним моделям, інформаційною базою яких є сукупності первинних даних.

Регресійна модель описує об'єктивно існуючі між явищами кореляційні зв'язки. У лінійному щодо параметрів рівнянні регресії індивідуальне значення модельованого показника y_h (де h — порядковий номер одиниці сукупності) записується так:

$$y_h = b_0 + \sum_{i=1}^{m} b_i x_i + e_h,$$

де b_i — *коефіцієнт регресії*, він показує, як в середньому змінюється y зі зміною x_i на одиницю її шкали вимірювання за незмінності решти включених у модель факторів і за інших рівних умов; b_0 — вільний член рівняння; економічного змісту, як правило, не має, лише окреслює область існування моделі; $e_h = y_h - Y_h$ — *залишкова величина.*

У регресійній моделі основне навантаження покладається на коефіцієнт регресії b_i, він розглядається як своєрідна міра "очищеного" впливу x_i на y і називається *ефектом впливу.* Значення коефіцієнтів регресії певною мірою залежать від складу введених у модель факторів. З розширенням ознакової множини моделі відбувається перерозподіл впливу попередньо введених факторів. Чим вагоміший вплив нововведеного фактора, тим помітніші зміни.

Процедура оцінювання параметрів регресійної моделі ґрунтується на методі найменших квадратів (МНК), алгоритми якого описані в математико-статистичній літературі і реалізовані в комп'ютерних програмах. Всесвітньо відомі статистичні пакети для комплексної обробки даних: *BMDP, SPSS, SAS, Statgraphics*. З 1995 р. світовим лідером на ринку статистичного програмного забезпечення визнається інтегрована система *Statistica* (версія 5.0), один з модулів якої — *Multiple Regression* — включає вичерпний набір засобів множинної регресії та оцінювання адекватності моделей реальним процесам [2].

Адекватність регресійної моделі означає здатність її правильно описати реальну структуру взаємозв'язків між ознаками x_i та y. Методологічною основою розв'язання проблеми адекватності є теоретичний, змістовний аналіз матеріальної природи явища й обґрунтування типу та структури моделі, яка описує механізм його формування. Практично з метою забезпечення адекватності моделі змістовний аналіз поєднується з формальними процедурами перевірки гіпотез відносно дотримання логіко-статистичних умов використання МНК. З-поміж цих умов зазначимо: достатнє число ступенів вільності варіації, однорідність сукупності, незалежність спостережень, інформативність включених і стабільність невключених у модель факторів, тип моделі. Математичною основою дотримання цих умов слугує ймовірний розподіл залишків e_h.

Використовуючи процедури модуля *Multiple Regression*, за умовними даними розглянемо можливості адаптації регресійних моделей до специфіки соціально-економічної інформації.

У моделях класичної регресії факторні ознаки x_i належать до метричної шкали вимірювання, виражаються числом і їхнє значення варіює в певних межах. У соціально-економічних дослідженнях часто стикаються з ситуацією, коли окремі властивості явищ нечислові, текстові (форма власності, професія тощо). Це ознаки номінальної шкали — шкали найменувань, градацій. Одиниці сукупності з однаковими градаціями утворюють групи (класи). При включенні такої ознаки в регресійну модель кожній k-й градації приписується число u_k, що має лише два значення (0; 1).

Оцифрування ґрунтується на дотриманні двох умов: повноти шкали і неперетинальності градацій. Повнота шкали градацій дає: $\sum u_k = f_k$, де f_k — частота k-ї градації. Для кожної з них середнє значення дорівнює частці $\overline{u} = \frac{f_k}{n} = d_k$. Умова неперетинальності виключає одночасну належність одиниці сукупності до двох градацій: $\sum u_k u_s = 0$, де k, s — градації ($k \neq s$).

Оскільки величина u_k є характеристикою розподілу сукупності, то надалі будемо її називати *структурною змінною*. Ця змінна розглядається як умовний код, що вказує на належність (1) чи неналежність (0) j-ї одиниці сукупності до k-ї градації. Для ознаки, що має p градацій $x_1, x_2, \ldots, x_p$, ставиться у відповідність $(p-1)$ величин $u_1, u_2, \ldots, u_{p-1}$. У регресійному аналізі до матриці ознакової множини моделі $X = [x_1, x_2, \ldots, x_m]$ додається матриця структурних змінних $U = [u_1, u_2, \ldots, u_{p-1}]$, а модель включає додаткові члени $a_1u_1 + a_2u_2 + \ldots + a_{p-1}u_{p-1}$. Параметри a_k оцінюються одночасно з коефіцієнтами регресії b_i при ознаках метричної шкали. Так, наприклад, у чорній металургії енерговитрати на одну тонну готового прокату y залежать від структури сталеплавильних процесів і способу виплавки сталі. Перший фактор x_1 характеризується часткою сталі, що розливається на машинах неперервного литва заготовок, другий x_2 — належить до номінальної шкали і має три градації: киснево-конверторний, електродуговий і мартенівський. В ознакову множину моделі фактор x_2 можна ввести двома структурними змінними:

Спосіб виплавки сталі	"21	"22
Киснево-конверторний	1	0
Електродуговий	0	1
Мартенівський	0	0

За такого варіанту оцифровування мартенівський спосіб виплавки сталі дістає еквіваленти (0, 0) і стає базою порівняння для двох інших. В регресійній моделі енерговитрат на тонну прокату $Y = a_0 + b_1x_1 + a_{21}u_{21} + a_{22}u_{22}$ параметр b_1 характеризує чистий ефект впливу структури сталеплавильного процесу, незалежно від способу виплавки сталі, a_{21} показує різницю в енерговитратах між киснево-конверторним і мартенівським способами виплавки сталі, незалежно від структури сталеплавильного процесу, a_{22} має таку ж саму інтерпретацію для електросталі.

Загальний вигляд регресійної моделі зі структурними змінними:

$$Y = a_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=q+1}^{m} \sum_{k=1}^{p-1} a_{rk} u_{rk}.$$

Коефіцієнти регресії вимірюють:

b_i — чистий, елімінований від взаємозв'язків в середині моделі, ефект впливу фактора x_i;

a_{rk} — вплив k-ї градації r-го фактора ($r \neq i$) на функцію y; алгебраїчно це різниця середніх значень функції y між k-ю градацією і градацією, взятою за базу порівняння.

Матриця первинних даних такої моделі складається з двох блоків: перший — блок факторних ознак метричної шкали обсягом (q*n), другий — блок структурних змінних для ознак номінальної шкали обсягом [($m - q$)*n]. В системі *Statistica* приписування текстовим ознакам числових еквівалентів здійснюється за опціями *Text Values Manager*, в якому реалізовано механізм подвійного запису ознак (*Text Value = Numeric*), і в процесі обробки даних можна переключатися з одного типу даних на інший.

Розглянута методика використання структурних змінних передбачає, що всі одиниці сукупності мають градації існуючої шкали. Якщо ця умова не виконується, то можна ввести додаткову групу для невизначених градацій. Не завжди виконується й умова неперетинальності груп — одна і та ж одиниця сукупності може одночасно належати до різних градацій. Скажімо, робітник має кілька професій, і для забезпечення умови неперетинальності його відносять до градації, яка відповідає основній професії. Аналогічна проблема виникає при обробці даних соціологічних обстежень, програмою яких передбачені питання-набори. Наприклад, респондент може вказати декілька джерел інформації про валютний ринок: радіо, телебачення, пресу. Кожна градація набору розглядається як альтернативна ознака і може самостійно включатися в модель [5].

Класична регресія передбачає також однорідність сукупності, тобто усі одиниці сукупності мають бути однотипними щодо комплексу умов існування, а властиві їм закономірності однаковими для всіх одиниць без винятку. За допомогою структурних змінних можна адаптувати регресійну модель до неоднорідної сукупності. Якщо неоднорідність проявляється розшаруванням сукупності на p ізольованих класів, то кожен клас розглядається як градація номінальної ознаки і тим одиницям, що належать до j-го класу, приписується структурна змінна $u_j = 1$, а тим, що не належать,— $u_j = 0$. Параметри за структурних змінних класів інтерпретуються так само як і при градаціях текстових ознак.

Специфіка моделювання процесів у неоднорідній сукупності зумовлена своєрідністю внутрішньокласової варіації і характером взаємозв'язків. Ефекти впливу одного і того ж фактора x_i на y за класами можуть істотно різнитися. Наприклад, у вугільній промисловості виділяються класи шахт за гірничо-геологічними умовами: потужністю та нахилом залягання пластів, їх газоносністю, глибиною розробки лав тощо. Кожному типові цих природних умов відповідає певна технологічна схема і певний рівень механізації виробничих процесів. Вплив механізації на трудомісткість, скажімо, очисних робіт залежить від класу шахти. Залежність сили впливу одного фактора від рівня іншого називається *взаємодією*. В неоднорідних сукупностях йдеться про взаємодію факторів і специфічних умов окремих класів. Для цього використовують змінні взаємодії $x_i u_j$, значення яких дорівнює добутку значень відповідних ознак.

Отже, при моделюванні взаємозв'язків в неоднорідних сукупностях ознакова множина моделі включає, окрім факторних ознак x_i, два типи інструментальних змінних: структурні змінні u_j, які відображують особливості класів, і змінні взаємодії $x_i u_j$, що характеризують особливості взаємозв'язків в окремих класах. За рахунок цих змінних модель регресійного аналізу розширюється:

$$Y = a_0 + \sum_{i=1}^{m} b_i x_i + \sum_{j=1}^{p} a_j u_j + \sum_i \sum_j c_{ij} x_i u_j ,$$

де b_i характеризує чистий ефект впливу i-го фактора в середньому за сукупністю; a_j показує відхилення середнього значення показника-функції в j-му класі від середнього його рівня в класі, взятому за базу порівняння; c_{ij} є різницею ефектів впливу i-го фактора в j-му класі порівняно з середнім впливом в цілому по сукупності.

Істотність параметрів a_j та c_{ij} свідчить про неоднорідність сукупності.

Модель такого типу Е. Маленво [4] назвав коваріаційною. В ній поєднуються регресія на факторних ознаках метричної шкали і модель дисперсійного аналізу міжкласових відмінностей. Вона уможливлює одночасну оцінку декількох рівнянь, і така оцінка може бути більш точною, ніж оцінки покласових регресій.

Аналітичні можливості моделі розглянемо на прикладі продуктивності праці робітників очисних вибоїв на шахтах з різними гірничо-геологічними умовами. Одні шахти мають пологі пласти (перший клас), інші — крутопадаючі (другий клас). Акцентуючи увагу на особливостях коваріаційної моделі, обмежимося одним фактором — швидкістю просування лави.

Модель має вигляд:

$$Y = a_0 + b_1 x_1 + a_1 u_1 + c_{11} x_1 u_1 .$$

Параметри її наведено в табл. 1. У верхній частині вікна *Regression Summary* подаються значення коефіцієнтів щільності зв'язку: множинної кореляції R, множинної детермінації $R1$ та Adjusted R1 (скоригований на число ступенів вільності); значення F-критерія та стандартної похибки s_e — *Std error*. Безпосередньо в таблиці наводяться параметри моделі: вільний член рівняння регресії b_0 — *Intercept*, β_i-коефіцієнти і коефіцієнти регресії b_i із стандартними похибками, значення t-критерію і фактичні рівні істотності *p-level*.

Коефіцієнт детермінації свідчить, що 92,1 % варіації продуктивності праці робітників очисних вибоїв (y) пояснюється класом шахт (u_1) і швидкістю просування лави (x_1). Адекватність моделі підтверджується значеннями F-критерію і *p-level*, істотність впливу факторів — характеристиками t-критерію. На шахтах з пологими пластами (клас 1) середня місячна продуктивність праці в очисних вибоях на 57,5 т вища порівняно з шахтами, що мають крутопадаючі пласти (клас 2). Зі збільшенням швидкості просування лави на 1 м продуктивність праці зростає в середньому на 2,23 т, на шахтах першого класу ефект впливу цього фактора на 1,15 т менший за середній.

Таблиця 1

Regression Summary for Dependent Variable: y
R = 0,96 R1 = 0,921 Adjusted R1 = 0,907
F(3,17) = 66,4 p < 0,000 Std.Error of estimate: 11,05

Con-tinue...	BETA	St. Err. of BETA	B	St. Err. of B	t(17)	p-level
Intercpt			43,525	15,078	2,89	0,010
x_1	1,31	0,296	2,227	0,504	4,42	0,000
u_1	0,80	0,286	57,500	20,484	2,81	0,012
$x_1 u_1$	-1,08	0,512	-1,153	0,548	-2,11	0,050

Теоретичний рівень продуктивності праці визначається так:

для шахт першого класу

$$Y = (43{,}525 + 57{,}5) + (2{,}227 - 1{,}153)x_1 ,$$

для шахт другого класу

$$Y = 43{,}525 + 2{,}227 x_1 .$$

Аналогічно здійснюється модельна специфікація за наявності трьох і більше класів. Наприклад, ознакова множина моделі включає дві факторні ознаки (x_1, x_2) та дві структурні змінні (u_1, u_2):

Структурні змінні		Специфікація моделі
u_1	u_2	
0	0	$Y = a_0 + b_1 x_1 + b_2 x_2$
1	0	$Y = (a_0 + a_1) + (b_1 + c_{11}) x_1 + b_2 x_2$
0	1	$Y = (a_0 + a_2) + b_1 x_1 + (b_2 + c_{22}) x_2$
1	1	$Y = (a_0 + a_1 + a_2) + (b_1 + c_{11}) x_1 + (b_2 + c_{22}) x_2$

За допомогою структурних змінних можна врахувати в моделі нетиповість певної групи одиниць, які класифікуються як аномальні. Належним до такої групи членам сукупності приписується $u_j = 1$.

Модель із структурними змінними і змінними взаємодії можна застосувати до комбінаційних групувань [5]. Традиційно для аналізу взаємозв'язків за даними комбінаційних групувань використовується модель дисперсійного аналізу *Anova / Manova*. Особливості застосування регресії до задач дисперсійного аналізу розглянемо на прикладі двофакторної класифікації (фактори *A* і *B*).

У регресійній моделі як і в дисперсійному аналізі, значення *i*-ї ознаки у *h*-ї одиниці сукупності, яка належить до *j*-ї групи, представляється сумою загальної середньої μ, ефекту кожного фактора $(a_i + b_j)$ та ефектів їх взаємодії $(ab)_{ij}$:

$$Y_{ijh} = \mu + a_i + b_j + (ab)_{ij} + e_{ijh},$$

де e_{ijh} — залишок; *i* — рівень фактора *A*; *j* — рівень фактора *B*.

Щоб забезпечити однозначність МНК-оцінок параметрів моделі формулюються додаткові обмеження:

$$\sum_i a_i = \sum_j b_j = 0,$$

$$\sum_i (ab)_{ij} = \sum_j (ab)_{ij} = 0.$$

На основі введених обмежень можна представити одні ефекти моделі як лінійну комбінацію інших і записати модель з мінімальною кількістю ефектів [1].

Наприклад, якщо за фактором *A* виділено три групи, а за фактором *B* — дві, то з урахуванням додаткових обмежень визначальними будуть параметри: μ, a_1, a_2, b_1, $(ab)_{11}$ і $(ab)_{21}$. Це мінімальна їх кількість, і модель з цими параметрами записується так:

$$Y = \mu + a_1 x_1 + a_2 x_2 + b_1 x_3 + (ab)_{11} x_4 + (ab)_{21} x_5.$$

Порядок формування файла первинних даних розглянемо на прикладі моделі тривалості перерви в роботі безробітних (*n* = 12). Фактор *A* — вікова група безробітних: A_1 — до 30 років; A_2 — від З0 до 50; A_3 — 50 років і старше. Фактор *B* — стать безробітного: B_1 — чоловіки; B_2 — жінки. У табл. 2 належність безробітного до відповідної підгрупи за цими факторами вказується подвійним індексом *y*, де / = 1, 2, 3;у = 1, 2; *y* — тривалість перерви в роботі, місяців.

Таблиця 2

ij	*y*	x_1	x_2	x_3	x_4	x_5
11	4	1	0	1	1	0
11	3	1	0	1	1	0
21	5	0	1	1	0	1
31	7	-1	-1	1	-1	-1
31	6	-1	-1	1	-1	-1
31	5	-1	-1	1	-1	-1
12	2	1	0	-1	-1	0
12	4	1	0	-1	-1	0
12	3	1	0	-1	-1	0
22	7	0	1	-1	0	-1
22	6	0	1	-1	0	-1
32	11	-1	-1	-1	1	1

Ознакова множина моделі *X* являє собою матрицю коефіцієнтів при відповідних ефектах впливу факторів та ефектах їхньої взаємодії. Так, x_1 відповідає ефектові a_1, тому $x_1 = 1$ для першої вікової групи, $x_1 = 0$ для другої вікової групи, а оскільки $a_1 + a_2 + a_3 = 0$, то для третьої вікової групи $x_1 = -1$. Аналогічно визначається вектор x_2, який відповідає ефектові a_2. Вектор x_3 належить до ефекту b_1. Прийнявши $x_3 = 1$ для чоловіків, маємо $x_3 = -1$ для жінок. Вектори x_4 та x_5 відносяться до ефектів взаємодії: $x_4 = x_1 x_3$; $x_5 = x_2 x_3$.

До сформованого таким чином файла первинних даних застосуємо процедури модуля *Multiple Regression*. Параметри моделі наведено в табл. 3. Істотними виявилися ефекти впливу першої групи фактора *A* (вік до З0 років) та першої градації фактора *B* (чоловіки), а також ефект взаємодії цих фактоів $(ab)_{11}$. Якщо середня тривалість в роботі по сукупності в цілому становить 5,8 міс, то у віковій групі до З0 років цей показник на 2,6 міс. менший; на 1 міс. менша за середню тривалість перерви в роботі у чоловіків.

При збільшенні кількості факторів оцінка ефектів впливу кожного з них і всіх можливих взаємодій за розглянутою методикою значно ускладнюється. У такому разі ефективною виявляється модель з адитивними ефектами. Адитивність означає незалежність впливу одного фактора від рівня іншого. Забезпечити її можна шляхом стандартизації комбінаційного групування, себто заміною частот емпіричного розподілу частотами певного стандартного розподілу. Зважені за частотами стандартного розподілу середні j-ї групи за i-м фактором Y_{ij} називаються стандартизованими, а відхилення цих середніх від загальної середньої $\bar{y}$ — стандартизованими (центрованими) ефектами $a_{ij} = Y_{ij} - \bar{y}$. Незсунені і з мінімальною дисперсією оцінки ефектів a_{ij} дає стандартизація групувань методом найменших квадратів [5].

Таблиця 3

Regression Summary for Dependent Variable: y R = 0,96 R1 = 0,922 Adjusted R1 = 0,857 F(5,6) = 14,22 p < 0,0028 Std. Error of estimate: 0,913						
Con-tinue…	BETA	St. Err. of BETA	B	St. Err. of B	t(6)	p-level
Intercpt, μ			5,833	0,291	20,02	1,01E-06
x_1	-0,962	0,141	-2,583	0,378	-6,84	0,000
x_2	-0,027	0,143	-0,083	0,435	-0,19	0,854
x_3	-0,432	0,126	-1	0,291	-3,43	0,014
x_4	0,448	0,135	1,25	0,378	3,31	0,016
x_5	0,078	0,136	0,25	0,435	0,57	0,586

У моделі МНК стандартизована середня Y_{ij} подається як регресія на структурних змінних u_{ij}:

$$Y_{ij} = \bar{y} + \sum_{i=1}^{m}\sum_{j=1}^{p_i} a_{ij} u_{ij},$$

де m — кількість факторів; p_i — кількість груп за i-м фактором.

Параметри моделі a_{ij} характеризують чистий, елімінований від взаємодії ефект впливу j-го рівня i-го фактора. Оцінювання параметрів моделі здійснюється розв'язуванням системи нормальних рівнянь, елементами якої є групові $\sum u_{ij}$ і підгрупові $\sum u_{ij} u_{rj}$ частоти. Щоб уникнути лінійної залежності рівнянь, ефекти однієї з груп за кожним фактором прирівнюються до нуля. Трансформована таким чином модель набуває вигляду:

$$Y_{ij} = b_0 + \sum_{i=1}^{m}\sum_{j=1}^{p_i-1} b_{ij} u_{ij},$$

де b_{ij} — різниця між груповими середніми j-ї групи і групи з нульовими ефектами (бази порівняння за i-м фактором).

Як і в загальній моделі МНК систему нормальних рівнянь такої моделі можна записати у матричному вигляді:

$$U'y = U'UB,$$

де $B = U'y(U'U)^{-1}$, $U'U$ — матриця структурних змінних.

Щоб визначити стандартизовані ефекти за i-м фактором a_{ij}, необхідно зафіксувати на середньому рівні значення інших факторів. Це досягається шляхом приписування всім структурним змінним, окрім змінної за i-м фактором, їхніх середніх значень, себто часток груп d_{ij}. Структурній змінній j-ї групи за i-м фактором приписується одиниця, іншим групам за цим фактором — нулі. У матричному вигляді розрахунок стандартизованих середніх записується у вигляді скалярного добутку вектора B на вектор коефіцієнтів D:

$$Y_{ij} = D'B.$$

Середня похибка стандартизованого ефекту оцінюється за формулою:

$$\mu_Y = s_e \sqrt{D'(U'U)^{-1} D},$$

де s_e — стандартна похибка.

Отже, схема стандартизації комбінаційних групувань МНК об'єднує два блоки. У першому визначаються параметри трансформованої моделі b_{ij}, у другому — вектор коефіцієнтів переходу від параметрів $Бц$ до стандартизованих ефектів a_{ij}.

Реалізацію цієї схеми розглянемо на прикладі професійної мобільності зареєстрованих безробітних. Як оцінку професійної мобільності використаємо частку безробітних, які виявили бажання пройти перенавчання для роботи за іншою професією.

У табл. 4 подано комбінаційний розподіл безробітних за віком (фактор A) та освітою (фактор B). Вікова структура представлена двома групами: A_x — до 3О років, A_2 — 3О років і старше; освіта — трьома рівнями: B_x — ПТУ, B_2 — середня спеціальна, B_3 — вища.

Таблиця 4

Вік, років	Рівень освіти			Разом	Направлено на перенавчання	Рівень професійної мобільності
	ПТУ	Середня спеціальна	Вища			
До 3О	150	40	10	200	60	0,30
Понад 3О	120	130	50	300	60	0,20
Разом	270	170	60	500	120	0,24
Направлено на перенавчання	53	43	24	120	X	X
Рівень професійної мобільності	0,20	0,25	0,40	0,24	X	0,24

Як видно з даних табл. 4, молодь (до 3О років) професійно мобільніша. Водночас простежується залежність професійної мобільності від рівня освіти: чим вищий рівень освіти, тим вища готовність набути нову професію. Очевидно, що ці два фактори взаємозв'язані. Стандартизація групування МНК передбачає перш за все ідентифікацію структурних змінних. Вилучивши за кожним фактором останню групу, матимемо одну структурну змінну за фактором *А* і дві — за фактором *В*. Рівняння регресії має вигляд:

$$Y = b_0 + b_{11}u_{11} + b_{21}u_{21} + b_{22}u_{22}.$$

Враховуючи, що $u_{ij}^2 = u_{ij}$, а $u_{ij}u_{ik} = 0$, де $j \neq k$, елементами матриці системи нормальнігх рівнянь $U'U$ та вектора $U'y$ будуть такі частоти:

$$U'U = \begin{vmatrix} n & \sum u_{11} & \sum u_{21} & \sum u_{22} \\ \sum u_{11} & \sum u_{11} & \sum u_{11}u_{21} & \sum u_{11}u_{22} \\ \sum u_{21} & \sum u_{11}u_{21} & \sum u_{21} & 0 \\ \sum u_{22} & \sum u_{11}u_{22} & 0 & \sum u_{22} \end{vmatrix} = \begin{vmatrix} 500 & 200 & 270 & 170 \\ 200 & 200 & 150 & 40 \\ 270 & 150 & 270 & 0 \\ 170 & 40 & 0 & 170 \end{vmatrix}$$

$$(U'y)' = \left(\sum y, \sum yu_{11}, \sum yu_{21}, \sum yu_{22}\right) =$$

$$= (120;\ 60;\ 53;\ 43).$$

Розв'язавши систему рівнянь, дістаємо параметри:

$$B' = (0{,}3470;\ 0{,}1557;\ -0{,}2642;\ -0{,}1577),$$

які оцінюють ефекти відповідних груп за /-м фактором відносно вилученої групи. На основі цих параметрів визначаються стандартизовані середні Y_{ij} та ефекти a_{ij}:

$$Y_{ij} = b_0 + b_{ij}u_{ij} + \sum_{1}^{m-1}\sum_{1}^{p_r} b_{rj}d_{rj},$$

де $a_{ij} = Y_{ij} - \bar{y}$; d_{rj} — частка j-ї групи за r-м фактором $(i \neq r)$.

Сформуємо вектори коефіцієнтів переходу від регресійної моделі до стандартизованих середніх Y_{ij} за даними табл. 4 ($m = 2$; $p_1 = 2$; $p_2 = 3$):

$$D_{11} = (1,\ 1,\ 0,\ d_{21},\ d_{22},\ d_{23})$$
$$D_{12} = (1,\ 0,\ 1,\ d_{21},\ d_{22},\ d_{23})$$
$$D_{21} = (1,\ d_{11},\ d_{12},\ 1,\ 0,\ 0)$$
$$D_{22} = (1,\ d_{11},\ d_{12},\ 0,\ 1,\ 0)$$
$$D_{23} = (1,\ d_{11},\ d_{12},\ 0,\ 0,\ 1).$$

Комбінаційний розподіл безробітних (табл. 4) характеризується частками: $d_{11} = 200 : 500 = 0{,}4$; $d_{12} = 0{,}6$; $d_{21} = 270 : 500 = 0{,}54$; $d_{22} = 170 : 500 = 0{,}34$; $d_{23} = 0{,}12$. Звідси стандартизована середня для першої групи за фактором *А* становить:

$$Y_{11} = 0{,}3470 + 0{,}1557 + 0 + (-0{,}2642)*0{,}54 +$$
$$+ (-0{,}1577)*0{,}34 + 0 = 0{,}306.$$

Відповідно центрований ефект цієї групи $a_{11} = 0{,}306 - 0{,}240 = +0{,}066$, тобто професійна мобільність молоді, незалежно від освіти, в середньому на 6,6 % вища за середній рівень. Аналогічно визначені стандартизовані середні і центровані ефекти для інших груп наведено в табл. 5.

Згідно з даними таблиці професійна мобільність осіб старшого віку на 8,8 % нижча за середній рівень. Щодо освіти, то незалежно від віку, найбільший ефект професійної мобільності дає вища освіти (+16,9%). Для випускників ПТУ, навпаки, характерний низький рівень професійної мобільності, ефект цієї групи становить (-9,5 %).

Таблиця 5

Номер групи, ij	11	12	21	22	23
Y_{ij}	0,306	0,152	0,145	0,251	0,409
a_{ij}	0,066	-0,088	- 0,095	0,011	0,169

Отже, за допомогою стандартизації ефекти впливу факторів, представлених ознаками різного типу (метричних, номінальних), приводяться до порівняльного виду. І це значно розширює аналітичні можливості регресійних моделей.

1. *Афифи А., Эйзен С.* Статистический анализ: Подход с использованием ЭВМ. Пер. с англ.— М: Мир, 1982.

2. *Боровиков В. П., Боровиков И. П.* Statistica® — Статистический анализ и обработка данных в среде Windows®.— М.: Издательский дом "Филинъ", 1998.

3. *Єріна А. М.* Статистичне моделювання та прогнозування: Навч. посібник.— К.: КНЕУ, 2001.

4. *Маленво Э.* Статистические методы эконометрии. Выпуск 1.— М: Статистика, 1975.

5. *Трофимов В. П.* Логическая структура статистичеких моделей.— М.: Финансы и статистика, 1985.

Yerina A. M.

CONJUNCTION OF THE MODELS OF REGRESSION AND VARIANCE ANALYSIS IN SOCIAL AND ECONOMIC INVESTIGATIONS

In this article the author is describing the analiticale possibilities of the method of extended regression and variants of its applications to the analysis of the non-homogeneous agregates of the sorted datas.